学校编码：10384　　分类号_____密级_____
学号：19020111152526　　UDC _____

厦门大学

硕 士 学 位 论 文

HILL 统计量估计极值指数时位置可变性研究

Research on position variability of HILL statistics estimated Extreme Value Index

刘坤宾

指导教师姓名：林涛 副教授
专 业 名 称：概率论与数理统计
论文提交日期：2014 年 4 月 10 日
论文答辩时间：2014 年 5 月 15 日
学位授予日期：2014 年 6 月 30 日

答辩委员会主席：王海斌教授
评 阅 人：杨静平教授、王芳副教授

2014 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行)》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人（签名)：刘坤宾

2014 年 5　月 20 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：刘坤宾

2014 年　5 月 20 日

摘 要

极值理论对样本尾部分布的极值指数的估计方法主要有两类：半参数方法和全参数方法，前者主要是基于分布尾部的 Hill 估计量，后者则主要基于广义帕累托分布。本文对两类 Hill 型估计量进行如下研究：由于 Hill 估计量对正极值指数的估计是位置可变的，本文对其位置平移后估计的变化和特点采用蒙特卡罗随机模拟方法进行模拟分析；另外，本文对位置不变的改进型 Hill 估计量的估计效果与 Pickands 估计、Hill 估计进行比较分析并得出的结论：针对 Frechet-Pareto 型分布样本，在一般情况下，位置不变的 Hill 估计量可作为估计极值指数的首选。

关键字：极值理论；Hill 型估计量；随机模拟

ABSTRACT

Methods for the estimation of Extreme Value Index of the tail distribution in Extreme Value Theory are of two main types: semi parameter method and parameter method, the former is the tail of a distribution based on Hill estimator, the other is based on the Generalized Pareto Distribution. In this paper, two kinds of Hill estimator type are reasearched as follows: as the Hill estimation for positive Extreme Value Index is variable for its position, we get the characteristics of estimates when its position shifted by Monte Carlo simulation method; this paper also reasearched the effect of Location-invariant Hill estimator compared with Pickands estimator and Hill estimator ,the conclusion is that ,in general, the Location-invariant Hill estimator can be used to estimate the Extreme Value Index of first choice based on the type of Frechet-Pareto samples.

Keywords: Extreme Value Theory ;Hill-type estimator ;Stochastic Simulation

目 录

Contents

第一章 绪论

1.1 研究背景

极值理论是概率论与数理统计方向的一个重要分支，极大值和极小值统称为极值，其极限分布问题即为极值理论，极值理论建立的极值分布模型只以分布的尾部数据为样本，它能够非常精准的描绘分布尾部的分位数，是一门用来预测异常现象或者小概率事件风险的模型技术，如今在气象、金融、保险、材料强度、洪水、地震等许多领域有广泛的应用。近几年来，极值理论的研究取得了很大进展，对机制理论感兴趣的，已由最初的概率论的研究人员及实际应用部门，发展到现在的主流统计学家，在各个领域都在不断地进行理论与应用的创新。

在极值估计中，对极值指数(Extreme Value Index)的估计构成了极值估计的主要内容，极值指数反映了尾部数据分布的变化速率，目前来说，极值理论对数据分布尾部指数的估计方法主要有两类：半参数方法和参数方法，前者主要是基于分布尾部的Hill估计量，后者则是基于广义Pareto分布。此外，许多学者对的估计提出了多种估计量，如Pickands估计、Moment估计、Hill估计、指数回归模型估计、基于GPD分布的POT估计法等。

1.2 本文研究内容与思路

本文对两类 Hill 型估计量进行如下研究：Hill 估计量对正极值指数的估计是位置可变的，本文对其位置平移后估计的特点采用蒙特卡罗随机模拟方法进行模拟分析；另外本文对位置不变的改进型 Hill 估计量的估计效果与 Pickands 估计、Hill 估计进行比较与分析并得出结论。

具体内容安排如下：

第二章主要介绍了极值理论的理论基础，重点介绍了其极值的三种分布类型、二阶条件和对极值指数(Extreme Value Index)的估计方法；

第三章针对ξ大于 0 时的 GEV 分布即 Frechet-pareto 族的 9 种分布，采用蒙特卡罗模拟给出 Hill 统计量以及平移后的 Hill 统计量随机模拟结果并得出结论：Hill 估计量平移以后的估计受一阶参数ξ影响大，受二阶参数ρ影响较小。

第四章为探讨位置不变 Hill 估计量的特点及估计效果，对 Frechet-pareto 族分布的 9 种不同类型分布，采用蒙特卡罗模拟给出位置不变 Hill 统计量、Hill 统计量、

Pickands 统计量估计极值指数随机模拟结果和均方误差 MSE 模拟结果，由此得出针对 Frechet-Pareto 型分布样本估计ξ，在一般情况下，位置不变的 Hill 估计量可作为首选。其中以上模拟与计算均使用 R 语言编程实现。

第二章 极值理论基础

2.1 极值理论概述

2.1.1 经典极值理论

极值理论是由Fisher 和Tippet(1920)首先提出的，随后Gennedenko(1940)进行了进一步的分析，最后Gumbel(1950)对极值的概率模型进行了标准化。这里我们主要介绍极值理论最基本的模型及其特征。

设 $X_1, X_2, ...X_n$ 是一列独立同分布的随机变量，共同的分布函数为F(x)，对自然数n，令

$$M_n = \max\{X_1, X_2, ..., X_n\}, \quad m_n = \min\{X_1, X_2, ..., X_n\}$$

分别表示这n个随机变量的最大值和最小值，则有

$$\Pr(M_n \le x) = \Pr(X_1 \le x, X_2 \le x, ..., X_n \le x) = F^n(x), x \in R,$$
$$\Pr(m_n \le x) = 1 - \Pr(m_n > x) = 1 - [1 - F(x)]^n, x \in R$$

这里R表示实数集合。由此可以知道，若知道总体分布F(x)，则无论x是否有限，当n趋近于无穷大时，M_n的分布极限只能是0或1，这样退化形式研究是没有意义的。在多数实践中，F往往是未知的，我们用n个随机变量最大值M_n的规范变换来了解最大值分布，得到以下定理：

定理 2.1.1（Fisher-Tippet）

设 $X_1, X_2, ...X_n$ 是独立同分布随机变量序列，如果存在常数列$\{a_n > 0\}$和$\{b_n\}$，使得对某一非退化分布函数H(x)：

$$\lim_{n\to\infty} \Pr\left(\frac{M_n - b_n}{a_n} \le x\right) = H(x), \ x \in R$$

则H必属于下列三种类型之一：

$$I: H_1(x) = \exp(-e^{-x}), -\infty < x < \infty$$
$$II: H_2(x;\alpha) = \begin{cases} 0, x \le 0 \\ \exp\{-x^{-\alpha}\}, x > 0 \end{cases} \alpha > 0$$
$$III: H_3(x;\alpha) = \begin{cases} 1, x > 0 \\ \exp\{-(-x)^{\alpha}\}, x \le 0 \end{cases} \alpha > 0$$

其中I型分布称为Gumbel分布，II型分布称为Frechet分布，III型分布称Weibull分布，

这三种分布统称为极值分布（extreme value distribution）。

若 $F(x)$ 的极值分布为 Frechet 分布型，那么有 $1-F(x)=x^{-\frac{1}{\alpha}}\ell_F(x)$，其中 $\ell_F(x)$ 为缓慢变换。即 $F(x)\in RV_{-1/\alpha}$。定义 $U(x)=F^{-1}(1-\frac{1}{x})=x^{\alpha}\ell_u(x)$，那么 $U(x)=x^{\alpha}\ell_u(x)$ 其中 $\ell_u(x)$ 为缓慢变换，即 $U(x)\in RV_{\alpha}$。

从模型角度来看，三种极值分布类型H1(x)、H2(x)和H3(x)完全不同，但从数学角度来看，它们之间却存在非常密切的关系。可以验证：X>0时有

$$\mathrm{X}\sim\mathrm{H}_2\Leftrightarrow\log\mathrm{X}^{\alpha}\sim\mathrm{H}_1\Leftrightarrow-\mathrm{X}^{-1}\sim\mathrm{H}_3$$

因此为方便起见，可以假定其中任意类型的极值分布。

容易求得三种极值分布的密度函数为：

$$\mathrm{I}: h_1(x)=e^{-x}H_1(x),-\infty<x<\infty$$
$$\mathrm{II}: h_2(x;\alpha)=\alpha x^{-(1+\alpha)}H_2(x;\alpha),x>0$$
$$\mathrm{III}: h_3(x;\alpha)=\alpha(-x)^{\alpha-1}H_3(x;\alpha),x\le 0$$

这三个密度函数都是单峰函数。

2.1.2 Frechet 分布型极值分布的二阶变换

若存在 $a(x)$ 使 $\frac{U(xu)-U(x)}{a(x)}\to\frac{u^{\alpha}-1}{\alpha},x\to+\infty$，称 $U(x)\in C_{-\alpha}(a(x))$。$\alpha$ 为极值指数，即一阶条件下的极值参数，其中 $a(x)=\alpha U(x)$。

若存在 $\rho<0$ 使

$$\lim_{x\to\infty}\frac{1}{a_2(x)}\left(\frac{U(xu)-U(x)}{\tilde{a}(x)}-h_{\alpha}(u)\right)=\tilde{c}h_{\alpha+\rho}(u)$$

成立，其中 $h_{\alpha}(u)=\frac{u^{\alpha}-1}{\alpha}$，那么我们称 ρ 为二阶极值指数参数。

对于分布函数 $F(x)$，若 $\exists b(x)\in RV_{\rho}$，且 $\rho<0$、$\alpha>0$，使

$$\lim_{t\to\infty}\frac{\frac{U(tx)}{U(t)}-x^{\alpha}}{b(x)}=x^{\alpha}\frac{u^{\rho}-1}{\rho}$$

成立，那么我们 $F(x)$ 满足二阶条件，ρ 为二阶极值指数参数，α 为一阶极值指数参

数，上式等价于使：

$$\lim_{x\to\infty}\frac{\log \ell_u(ux)-\log \ell_u(x)}{b(x)}=h_\rho(u)$$

即：

$$\lim_{x\to\infty}\frac{\log U(ux)-\log U(x)-\alpha\log x}{b(x)}=\frac{x^\rho-1}{\rho}$$

成立。可以得到，上式的二阶条件等价于$\exists c$ 使：

$$\log \ell(x)=c+dx^\rho+o(x^\rho)$$

成立，其中，c，d 都是常数。更多的关于二阶条件的内容和例子可以参阅 de Haan 和 Ferreira 的文章。下表 2.1 给出了常见 Frechet 分布型二阶参数。

表 2.1 几种 Frechet 分布的一阶和二阶参数

分布	1-F(x)	一阶 EVI	$b(x)$	二阶 ρ
Pareto(α)	$x^{-\alpha}$ $x>1;\alpha>0$	$\frac{1}{\alpha}$	0	0
GP(σ,γ)	$(1+\frac{\gamma x}{\sigma})^{-1/\gamma}$ $x>0,\sigma,\gamma>0$	γ	$b(x)=\gamma x^{-\gamma}$	$\rho=-\gamma$
Burr(η,τ,λ) Type XII	$(\frac{\eta}{\eta+x^\tau})^\lambda$ $x>0;\eta,\tau,\lambda>0$	$\frac{1}{\lambda\tau}$	$\frac{1}{\lambda\tau}x^{-1/\lambda}$	$\rho=-\frac{1}{\lambda}$
Fréchet(α)	$1-e^{-x^{-\alpha}}$ $x>0;\alpha>0$	$\frac{1}{\alpha}$	$\frac{1}{2\alpha x}$	$\rho=-1$
Burr(η,τ,λ) Type III	$1-(\frac{\eta}{\eta+x^{-\tau}})^\lambda$ $x>0;\eta,\tau,\lambda>0$	$\frac{1}{\tau}$	$\frac{1}{2\tau x}(\frac{1}{\lambda}+1)$	$\rho=-1$

2.1.3 广义极值分布

如果引进位置参μ数和刻度参数σ，则三种极值分布形式为：

$$\mathrm{I}: H_1(x;\mu,\sigma)=\exp\{-e^{-\frac{x-\mu}{\sigma}}\},-\infty<x<\infty$$

$$\mathrm{II}: H_2(x;\mu,\sigma,\alpha)=\begin{cases}0, x\le\mu\\ \exp\{-(\frac{x-\mu}{\sigma})^{-\alpha}\}, x>\mu\end{cases}\quad\alpha>0$$

$$\mathrm{III}: H_3(x;\mu,\sigma,\alpha)=\begin{cases}1, x>\mu\\ \exp\{-(-\frac{x-\mu}{\sigma})^{\alpha}\}, x\le\mu\end{cases}\quad\alpha>0$$

这三种分布可以用统一的形式来表示：

$$H(x;\mu,\sigma,\xi)=\exp\{-(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\},1+\xi(x-\mu)/\sigma>0$$

其中，μ，$\xi\in R$，$\sigma>0$，我们称H为广义极值分布(generalized extreme value distribution)简称GEV分布，当μ=0，σ=1时，称之为GEV标准形式。这个模型具有三个参数：位置参数μ，刻度参数σ，ξ为形状参数。参数ξ称为分布H的极值指数(也称为尾指数)，当ξ大于0时，由于它的尾部像幂分布一样衰减，称这种分布为厚尾分布，H对应称为Frechet族，Pareto分布，Burr分布，学生化t分布，对数Gamma分布都属于这种Frechet族。当ξ<0时，H对应着Weibull分布，均匀分布，逆Pareto分布和逆Burr分布就是这一类。当ξ= 0时，由于它的尾部像指数分布一样衰减，称之为Gumbel族，像常见的指数分布，正态分布，对数正态密度，Gamma分布都在其中。这样统一以后有利于简化统计分析。这样表示以后定理2.1.1可以写成：

设$X_1,X_2,...X_n$是独立同分布随机变量序列，如果存在常数列$\{a_n>0\}$和$\{b_n\}$，使得对某一非退化分布函数H(x):

$$\lim_{n\to\infty}\mathrm{P\ r}(\frac{M_n-b_n}{a_n}\le \mathrm{x})=\mathrm{H}(\mathrm{x}),\ \mathrm{x}\in\mathrm{R}$$

成立，则称H是定义在$\{x:1+\xi\frac{x-\mu}{\sigma}>0\}$上的GEV分布。此时相应的密度函数为：

$$h(x;\mu,\sigma,\xi)=\frac{1}{\sigma}H(x;\mu,\sigma,\xi)(1+\xi\frac{x-\mu}{\sigma})^{-(1+1/\xi)},1+\xi\frac{x-\mu}{\sigma}>0$$

如图2.1表示标准GEV分布在ξ=0.3，ξ=0，ξ=-0.3时的密度图像：

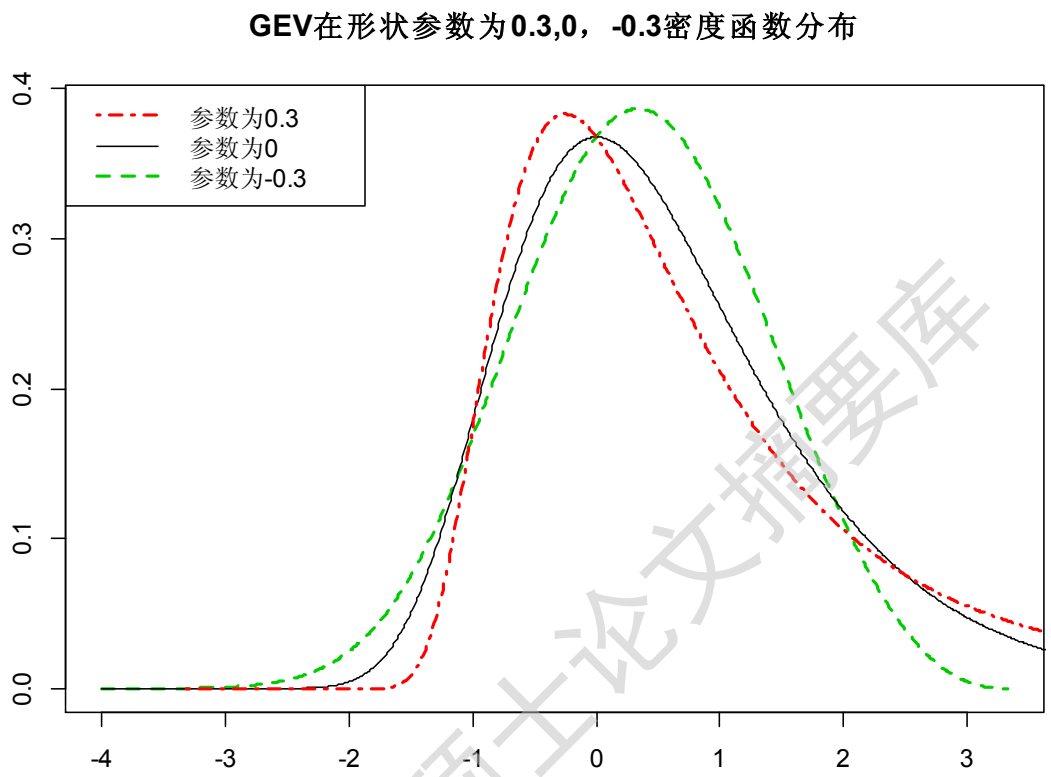

图2.1　标准GEV分布在ξ=0.3，ξ=0，ξ=-0.3 时的密度函数图

2.2　广义 Pareto 型分布

用 GEV 模型对实际数据建模时，一般是将观测值序列分成长度为 m 的 k 组，在每组中取出一个最大值，记为 M, 那么有 M1,M2,…Mk 就是每组的最大值数据，根据上节定理，只要 m 足够大，M1,M2,…Mk 就可以近视看成来自 GEV 分布的一个独立同分布观测。当用广义极值分布建立模型时，只能从 m 个数据中选取一个最大值进行建模，这样导致了对数据所包含信息的巨大浪费，又由于人们在实践中发现不只是数据的最大值才是极值，若有两个或两个以上的大值处于同一个长度为 m 的数据区间内，则只能选出一个大值来，而浪费了另一个或更多的大值，因此用最大值模型是低效的。但如果选定一个较大的阈值，那么超过这个阈值的所有数据都可以认为是极值，这样就可以充分利用极值数据所提供的信息。下面将介绍的广义 Pareto 分布就是反映这些超过阈值的极值数据统计特征的分布形式。

我们首先给出超阈值量分布：

设 X1,…,Xn 是独立同分布的随机变量序列，分布函数 F(x) 支撑的上端点为 x^*，选取一个阈值 u，称观测值大于阈值 u 的为超阈值 (peaks over threshold 简记为 POT)，记称 X_i-u 为超阈值量(excess)，则得到

$$F_u(x)=\Pr(X-u\le x\mid X>u)=\frac{F(x+u)-F(u)}{1-F(u)},\ x\ge 0$$

称 Fu(x)为随机变量 X 的超过阈值 u 的超出量的分布函数，简称超出量分布，对应的密度函数为：

$$f_u(x)=\frac{f(x+u)}{\bar{F}(u)},x\ge 0$$

称

$$e(u)=E(X-u\mid X>u)$$

为随机变量 X 的平均超出量函数 (mean excess function of X)。在实际应用中，F 往往是未知的，这就要考虑超出量的极限分布。

Rickands(1975)证明了如下定理：

定理 2.2.1 若存在常数 $a_u>0$ 和 b_u，使当 $u\to x^+=\sup\{x:0<F(x)<1\}$ 时，$Fu(a_u+b_u)$ 有连续极限分布，则：

$$\lim_{u\to x^+}\sup_{0\le x\le x^+-u}|F_u(x)-G(x;\sigma,\xi)|=0$$

对某个 ξ 与 $\sigma(u)$ (与 u 有关)成立。此定理说明在一般情况下，GPD 可作为超阈值量分布的近似分布。

如果随机变量 X 的分布函数为：

$$G(x;u,\sigma,\xi)=1-(1+\xi\frac{x-u}{\sigma})^{-1/\xi},x\ge u,\ 1+\xi(x-u)/\sigma>0$$

则称 X 服从广义 Pareto 分布 (generalized Pareto distribution)，简记为 GPD。可以求出，广义 Pareto 分布的密度函数为：

$$g(x;u,\sigma,\xi)=\frac{1}{\sigma}(1+\xi\frac{x-u}{\sigma})^{-\frac{1}{\xi}-1},x\ge u,1+\xi(x-u)/\sigma>0$$

我们可以看出 GPD 与上节所述之间的关系：若最大值 Mn 近似服从 GEV 分布，则在 X > u 的条件下，超出量 Y=X−u 和超阈值 X 近似服从 GPD，且与 GEV 分布有

相同的形状参数ξ，该定理的统计意义在于，我们可以用 GPD 来拟合超出量或超阈值，从而利用超出量分布或超阈值分布求总体分布函数。下图 2.2 表示 GPD 在u=0，σ=1 时，ξ取-0.4，0, 0.4 的分布函数图：

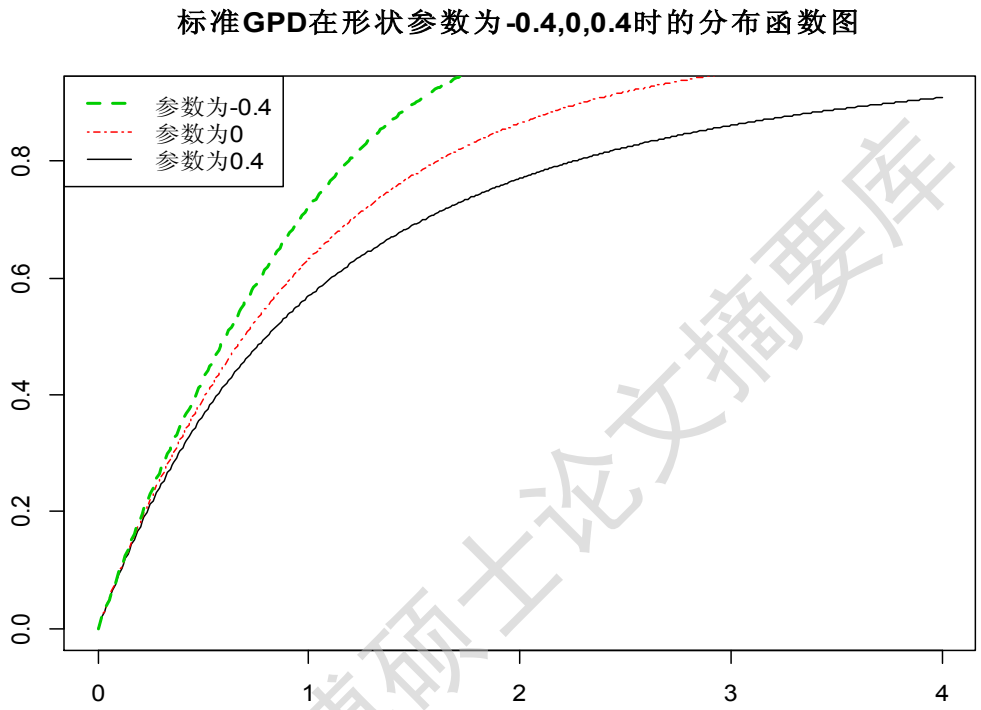

图2.2　标准GPD分布在ξ=-0.4，ξ=0，ξ=0.4时的分布函数图

2.3　极值分布的统计推断

我们假定$X_1, X_2, \cdots, X_n$是独立同分布的随机变量，$X_{1,n} \le X_{2,n} \le \ldots \le X_{n,n}$为其次序统计量。

极大似然估计法，GPD 分布的对数似然函数为：

$$l(\mu,\sigma,\xi)=\begin{cases}-n\log\sigma-(1+1/\xi)\sum_{i=1}^{n}\log\left[1+\xi(\frac{x_i-\mu}{\sigma})\right], \xi\neq 0\\ -n\log\sigma-\frac{1}{\sigma}\sum_{i=1}^{n}(x_i-\mu), \xi=0\end{cases}$$

令偏导数等于 0，得到方程：